CLAIMS

1. A system for distributing connections from clients on an external network to a plurality of servers on an internal network, the system comprising:

    a client interface to the external network the client interface being operative
5   to receive and send packets to and from a remote client;

    a server interface to the internal network, the server interface being operative to receive and send packets to and from a plurality of servers, the plurality of servers being operative to establish a connection with the remote client;

    a plurality of predicted responsiveness indicators, each of the plurality of
10 predicted response indicators being associated with each of the plurality of servers, the predicted responsiveness indicators being operative to predict the response time of each of the plurality of servers, the predicted responsiveness indicators also being stored within the system in a manner that the predicted responsiveness indicators may be accessed; and

15 a predicted responsiveness comparator which is operative to access and compare the predicted responsiveness indicators and to determine which servers from among the plurality of servers is associated with a predicted responsiveness indicator which measures a best response time, the predicted responsiveness comparator being further operative to select a pointer to a server which has a
20 predicted responsiveness that is the best predicted responsiveness among the predicted responsiveness of the plurality of servers.

    whereby the server which has a predicted responsiveness which is the best predicted responsiveness is selected to handle the next connection from a client. .

2. A system as recited in claim 1, wherein the predicted responsiveness
25 indicators are periodically updated.

3. A system as recited in claim 1, wherein the predicted responsiveness indicators include the number of connections to each of the plurality of servers.

4. A system as recited in claim 1, wherein the predicted responsiveness indicators include the number of connections to each of the plurality of servers.

5. A system as recited in claim 1, wherein the predicted responsiveness indicators include the predicted response time of each of the plurality of servers.

6. A system as recited in claim 1, wherein the predicted responsiveness indicators include the predicted response time of each of the plurality of servers and wherein the predicted response time is calculated according to the formula:

$$PR = NC*R - f(t_{current} - t_{age\ stamp})$$

where PR is the predicted response time, NC is the number of connections to the physical machine, R is a measured response time, and $f(t_{current} - t_{age\ stamp})$ is an aging function which is a function of the difference in time from the current time and a time recorded in an age stamp when the last response time was measured.

7. A system as recited in claim 6 wherein the aging function decreases the predicted response time as the time since the last predicted response increases.

8. A system as recited in claim 6 wherein R is a function of a plurality of measured response times.

9. A system as recited in claim 6 wherein R changes with each new measured response time, but is constrained to change by less than a certain percentage as a result of the latest measured response time.

10. A system as recited in claim 9 wherein the percentage is less than about 10%.

11. A system as recited in claim 9 wherein the percentage is less than about 25%.

12. A system as recited in claim 9 wherein the percentage is less than about 50%.

5   13. A method of distributing incoming packets among a plurality of physical machines comprising:

intercepting an incoming packet and determining a packet source IP address and a packet destination IP address for the packet from a header of the packet;

10   locating a connection data structure having entries which match the packet source IP address and the packet destination IP address;

using a connection data structure to determine a selected physical machine to which the packet is to be distributed;

replacing the packet destination IP address with an IP address of the selected physical machine; and

15   adjusting a checksum in the header of the packet to conform to the replaced packet destination IP address

whereby an incoming packet is distributed to a physical machine according to a selected physical machine determined by a connection data structure.

14. A method as recited in claim 13 wherein locating a connection data
20   structure having entries which match the packet source IP address and the packet destination IP address includes determining whether a predefined connection data structure exists and creating a new connection data structure if no a predefined connection data structure exists.

15. A method as recited in claim 14 wherein creating a new connection data structure includes determining which of the plurality of physical machines is best suited to handle the incoming packet.

16. A method as recited in claim 15 wherein determining which of the plurality of physical machines is best suited to handle the incoming packet includes using a foreign physical association data structure to determine whether the packet source address corresponds to a foreign entity which previously connected to a previously connected physical machine and distributing the incoming packet to the previously connected physical machine.

17. A method as recited in claim 15 wherein determining which of the plurality of physical machines is best suited to handle the incoming packet includes using a foreign physical association data structure to determine whether the packet source address corresponds to a foreign entity which previously connected to a previously connected physical machine and distributing the incoming packet to the previously connected physical machine if a maximum physical machine association time interval has not been exceeded.

18. A method as recited in claim 15 wherein determining which of the plurality of physical machines is best suited to handle the incoming packet includes determining which physical machine has the best predicted responsiveness.

19. A method as recited in claim 18 wherein determining which physical machine has the best predicted responsiveness includes determining which physical machine has the fewest number of connections.

20. A method as recited in claim 18 wherein determining which physical machine has the best predicted responsiveness includes determining which physical machine has the fewest number of connections relative to an individual weighting factor assigned to each physical machine.

21. A method as recited in claim 18 wherein determining which physical machine has the best predicted responsiveness includes determining which physical machine has the best predicted response time based at least in part on a previously measured response time.